(54) Title: TEXT SEGMENTATION AND TOPIC ANNOTATION FOR DOCUMENT STRUCTURING
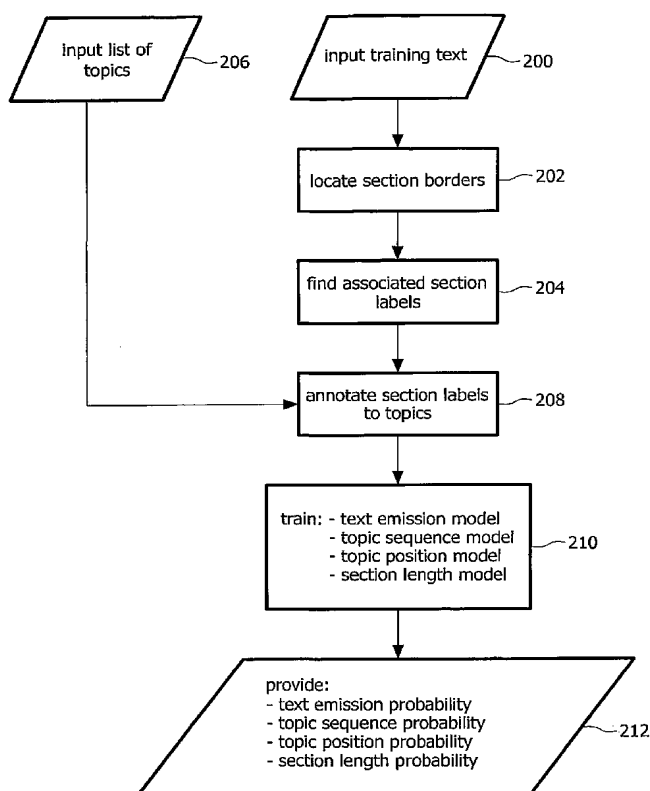
(57) Abstract: The invention relates to a method,
a computer program product and a computer system
for structuring an unstructured text by making use of
statistical models trained on annotated training data.
Each section of text in which the text is segmented is
further assigned to a topic which is associated to a set
of labels. The statistical models for the segmentation
of the text and for the assignment of a topic and its
associated labels to a section of text explicitly accounts
for: correlations between a section of text and a topic,
a topic transition between sections, a topic position
within the document and a (topic-dependent) section
length. Hence structural information of the training
data is exploited in order to perform segmentation and
annotation of unknown text.

PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report:
20 July 2006

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.